Introducing the *q-Theil* index

M. Ausloos[*]

GRAPES, ULG, B5a, Sart Tilman, B-4000 Liege, Euroland

J. Miśkiewicz[†]

Institute of Theoretical Physics, Wrocław University, pl. M. Borna 9, 50-204 Wrocław, Poland

Starting from the idea of Tsallis on non-extensive statistical mechanics and the *q-entropy* notion, we recall the Theil index Th and transform it into the Th_q index. Both indices can be used to map onto themselves any time series in a non linear way. We develop an application of the Th_q to the GDP evolution of 20 rich countries in the time interval [1950 - 2003] and search for a proof of globalization of their economies. First we calculate the distances between the "new" time series and to their mean, from which such data simple networks are constructed. We emphasize that it is useful to, and we do, take into account different time "parameters": (i) the moving average time window for the raw time series to calculate the Th_q index; (ii) the moving average time window for calculating the time series distances; (iii) a correlation time lag. This allows us to deduce optimal conditions to measure the features of the network, i.e. the appearance in 1970 of a globalization process in the economy of such countries and the present beginning of deviations. The q value hereby used is that which measures the overall data distribution and is equal to 1.8125.

Keywords: econophysics, time series analysis, entropy

1. INTRODUCTION

Since the fundamental work of Boltzmann [1] the entropy concept has been developed and applied to a range of subjects going from elementary termodynamics and statistical mechanics through quantum physics e.g. [2] information theory [3] up to applications in biology e.g. [4, 5, 6] and economy [7, 8, 9]. Recently Tsallis and many others in his path have shaken up the usual considerations on the entropy concept, in particular within Shannon information theory.

In fact, complex nonequilibrium systems can be often described by a superstatistics, which result of a superposition of two statistics associated with two different time scales [7, 10, 11, 12, 13, 14]. The methods of extracting superstatistics parameters from time series are discussed in [15]. In that line of thought, a special attention can be paid to the *entropy of a time series*.

On the other hand, the Theil [16] index is often used in economy and finance. It is defined through

$$Th(x;N) = \frac{1}{N}\sum_{i=t}^{N}\left(\frac{x_i}{\langle x_i \rangle}\ln\frac{x_i}{\langle x_i \rangle}\right) \tag{1}$$

where the average $\langle x \rangle$ is made over the ensemble of points N of the population of size N. It looks like the Shannon entropy but was invented to consider the event values themselves, in particular the income x_i of agent i in a population of N agents, rather than their probability of occurrence. One peculiarity is that it measures the individual's share of income relative to the mean income $\langle x_i \rangle$ of the population. With reference to information theory, Theil's measure is a difference between its maximum entropy and its present entropy at that time. Thus from the Theil index one can look at correlations between data sets, distances, hierarchies, and other usual features, through various techniques of data analysis, like those resulting after network constructions.

An interesting development is to consider that the x_i quantity in Eq.(1) is time dependent. Thus one can generalize the Theil index in order to remap in a nonlinear way a time series $x(t)$ into a $Th(t)$, as done in Sect. 2 which recalls considerations outlined in [17]. Moreover in the spirit of non-extensive statistical entropy, following Tsallis considerations, it can be imagined to propose the *q-Theil* index, as done in Sect. 2. The first application is here below made to macroeconomy time series, in particular to the GDP of the richest countries. Following up on studies of correlations between GDPs of rich countries [17, 18, 19, 20, 21, 22, 23, 24, 25], we have analyzed web-downloaded data on GDP, used as individual wealth signatures of a country economical state ("status"). We have calculated the fluctuations of the GDP and looked for correlations, and "distances", as reported in Sect. 3.

Usually, a system is represented by a network, nodes being scalar agents, here the countries, while links are weights, i.e. here measures of distances between two $Th(t)$ representing GDP fluctuation correlations between two countries. In order to extract structures from the networks, we have averaged the time correlations in different windows. This allows more robustness in the subsequent networks properties and reveals evolving *statistical distances*. In line with our previous works [17, 18, 19, 20] we have examined three different network constructions. A discussion on economy globalization follows with a conclusion in Sect. 4. It is found that such a measure of *collective habits* does fit the usual expectations defined by politicians or economists, i.e. common factors are to be searched for.

---

[*]Electronic address: marcel.ausloos@ulg.ac.be
[†]Electronic address: jamis@ift.uni.wroc.pl

2. THEIL INDEX AND TSALLIS ENTROPY

The original definition of the Theil index, see Eq.(1), allows for a peculiar mapping of a 1-D data set, like a time series, i.e. The Theil index nonlinearly maps the "original" time series $A(t)$ into a new one through

$$Th_A(t,T_1) = \frac{1}{T_1} \sum_{i=t}^{t+T_1} \left(\frac{A_i}{\langle A \rangle_{(t,T_1)}} \ln \frac{A_i}{\langle A \rangle_{(t,T_1)}} \right) \qquad (2)$$

where the average $\langle A \rangle_{(t,T_1)}$ is made over the ensemble of points j in a time window of size T_1, placed between t and $t+T_1$:

$$\langle A \rangle_{(t,T_1)} = \frac{1}{T_1} \sum_{j=t}^{t+T_1} A_j. \qquad (3)$$

Thus the Theil index is calculated for the interval $[t, t+T_1]$. Applications of the Theil index notion can be found in other papers [17, 20], where the Theil index was applied to measure the economy globalization process.

In order to connect with Tsallis non extensive statistics we introduce the *q-Theil* index Th_q^A for a time series $A(t)$

$$Th_q^A(t,T_1) = \frac{1 - \sum_{i=t}^{t+T_1} \left((A_i/\langle A_i \rangle_{(t,T_1)})^q \right)}{q-1} \qquad (4)$$

in the interval $[t, t+T_1]$. Eq.(4) corresponds to Eq.(2) when $q \to 1$.

In order to compare time series, their distance can be immediately introduced. Moreover the mean and standard deviations (std) of an ensemble of such distances can be used in further considerations. The distance between two time series (here the Theil-mapped time series) is hereby defined as the absolute value of the difference between mean values in the interval $[t, t+T_2]$. Moreover the elements of the time series can be taken at equal times or with the time lag τ, a possibility which we take also into consideration for generality purposes. Thus we define

$$d_{Th_q}(A,B)_{(t,T_1,T_2,\tau)} = |\langle Th_q^A(t,T_1) - Th_q^B(t+\tau,T_1) \rangle_{(t,T_2)}|. \quad (5)$$

In Eq.(5) the mean value denoted by brackets, $\langle ... \rangle$, is defined as in Eq.(3).

As a result we have three different time parameters:

1. the T_1 time window while calculating the Th_q index,

2. the time lag τ, and

3. the correlation window T_2.

Note: in the analysis both time windows (Th_q and correlation) are used congruently so the the total size of the time window is equal to the sum of the Th_q and correlation time windows. Therefore the number of the generated networks is equal to the time series length minus the total time window size.

3. MACROECONOMY INDEX INPUT AND NETWORK CONSTRUCTION

3.1. GDP data

GDP data sets of most rich OECD countries were used, i.e. 20 countries Austria, Belgium, Canada, Denmark, Finland, France, Greece, Ireland, Italy, Japan, the Netherlands, Norway, Portugal, Spain, Sweden, Switzerland, Turkey, U.K., U.S.A, and Germany, allowing for a linear superposition of the data before the reunification in 1991 in the latter case; an All country is also invented as in previous works [17, 18, 19, 20].[1] Thus $N = 21$. The data starts in 1950 and finish in 2003, so there are 54 data points in every time series.

3.2. Networks

The distance matrices obtained from Eq.(5) are analysed by constructing three network structures and analysing statistical properties of the distances between nodes. The following networks are considered: unidirectional minimal length path (UMLP), bidirectional minimal length path (BMLP) and locally minimal spanning tree (LMST). The algorithms generating the mentioned networks are:

UMLP The network begins with an arbitrary chosen country, – here the All country, then the closest neighbouring country is attached and become the end of the network. The next country closest to the end of the network is searched and attached. The process continued until all countries are attached to the network.

BMLP The network begins with the pair of countries with the smallest distance between them. Then the country closest to the ends of the network are searched and those with shorter distance attached to the appropriate end. The algorithm is continued until all countries become nodes of the network.

LMST The root of the network is the pair of closest neighbouring countries. Then the country closest to any node is searched and attached. The algorithm is continued until all countries are attached to the network.

Notice that in the UMLP construction, All is at the beginning of the chain, while in the other two constructions, All is treated as a "normal" country. The BMLP and LMST network seeds are the appropriate pairs of the closest countries according to the appropriate distance matrix. The first two networks are linear, and essentially robust against a "perturbation", like removing or adding a country or in the case of a regrettable mathematical error, since they are based on a measure relative to a statistical mean, while the LMST is obviously a tree,

---

[1] The set deviates somewhat from previous works [17, 19, 20] since there is neither Iceland nor Luxembourg but there is Turkey in the present paper.

rather compact when only 21 data points, thus with very few branching levels, are involved in the construction. It is known that such a tree is far from robust.

4. RESULTS

First let us report that a q value of the considered data set must be given. It could be let as a free parameter and one could find some optimal value according to some criterion, or a few criteria. Here below, i.e. for the GDP time series in 1950-2003 for the countries defined in Sec.3.1 we have calculated the q value for the following considerations by the maximum likelihood estimator [26], as for Tsallis entropy, and found $q = 1.8315$, hereby used to calculate distances through (Eq.(4)) and Eq.(5). It is fair to recall that Borges in [27] calculated the (Tsallis entropy) q value for GDP of USA, Brasil, Germany and UK. He found a q value varying from 1.4 (UK) up to 2.1 (Brasil), and or USA, $q=1.7$.

4.1. *q-Theil* distance statistics

In our analysis UMLP, BMLP and LMST networks were constructed for all time windows ranging from $T_1 = 5$ yrs , $T_2 = 1$ y moving along the time axis by a one year step. Eleven time lag values were considered: $\tau \in [0, 1, \ldots, 10]$. The T_1, T_2 and τ parameters statisfy the inequality $T_2 + T_3 + \tau \leq 54$ yrs, so the number of generated networks (*Net*) depends on the time window sizes and is equal to $N_{Net} = 54 - T_1 - T_2 - \tau$, for a given triplet, - times 3, due to the type of network considered. In total this is a huge number of networks. Therefore some cases are to be extracted for the present report.[2] Different presentations can be made, in a three dimensional time coordinate space. We propose a vizualisation of the data through a spectrogram method, using for the x and y axis the time window T_2 and T_1 respectively for a given τ. The data values are represented by a colored pixel in a convenient order.[3] The results of calculations of the mean but also values of the corresponding standard deviations are here below presented.

The mean value and standard deviation of the distances between nodes as a function of the T_1 and T_2 are presented in Figs. 1 - 9 for the time lag $\tau = 0, 5, 10$ yrs. The largest value of the mean distance, the minimum mean distance, the maximum and minimum standard deviations as a function of the time windows T_1, T_2 and time lags are presented in Table I.

It can be first generally observed that the mean distance between countries and the corresponding standard deviation are the biggest for UMLP networks and the smallest for LMST networks. It is also worth noticing that the mean distance depends on the time lag value. If the time lag is large the mean distance is large as well. The maximum of the mean distance
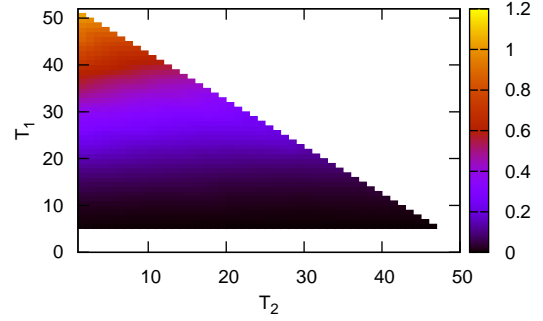

FIG. 1: Mean distance between countries as obtained in the case of the Th_q mapping and the UMLP network construction. The distance is averaged over the network links and the time. The values of T_1 and T_2 are given on the vertical and horizontal axis respectively. Time lag value: $\tau = 0$ y. The color scale in use is indicated. A few numerical values are given in Table I.
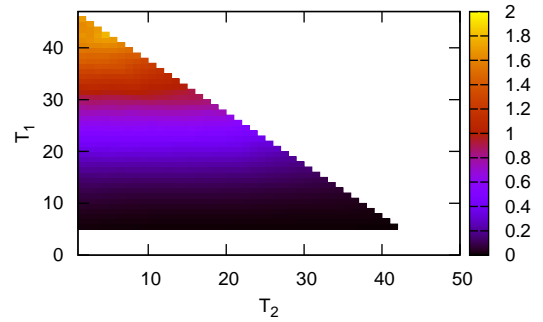

FIG. 2: Mean distance between countries deduced from Th_q for UMLP networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 5$ yrs. The color scale in use is indicated. A few numerical values are given in Table I.

occurs for the longest T_1 and the shortest T_2 windows sizes. The minimum mean distance is found with the oposite combination of the time windows sizes, i.e. small T_1 and large T_2. The standard deviations increase with the time lag and are the largest ones in the case of the longest considered time lag.

$q = 1.8315$		mean								std							
Network	τ (y)	max	T_1	T_2	N_{Net}	min	T_1	T_2	N_{Net}	max	T_1	T_2	N_{Net}	min	T_1	T_2	N_{Net}
UMLP	0	1.110	52	1	1	$18 \cdot 10^{-4}$	5	48	1	1.549	52	1	1	$29 \cdot 10^{-4}$	5	48	1
	5	1.956	43	5	1	$37 \cdot 10^{-4}$	5	37	7	2.414	43	5	1	$45 \cdot 10^{-4}$	5	43	1
	10	2.984	41	1	2	$61 \cdot 10^{-4}$	5	35	4	3.151	41	2	1	$72 \cdot 10^{-4}$	5	38	1
BMLP	0	0.8151	52	1	1	$15 \cdot 10^{-4}$	5	47	2	2.7829	51	2	17	$24 \cdot 10^{-4}$	5	48	1
	5	2.0014	46	2	1	$28 \cdot 10^{-5}$	5	41	3	2.0473	46	2	5	$41 \cdot 10^{-4}$	5	41	3
	10	2.7829	40	2	2	$56 \cdot 10^{-4}$	5	37	2	3.0620	40	2	2	$69 \cdot 10^{-4}$	5	37	2
LMST	0	0.8151	52	1	1	$15 \cdot 10^{-4}$	5	47	2	0.7671	51	2	1	$24 \cdot 10^{-4}$	5	48	1
	5	1.9609	45	2	2	$28 \cdot 10^{-4}$	5	43	1	1.8710	45	2	2	$43 \cdot 10^{-4}$	5	39	5
	10	3.2012	40	2	2	$52 \cdot 10^{-4}$	5	38	1	3.2334	41	2	1	$68 \cdot 10^{-4}$	5	38	1

TABLE I: The maximum mean distance, the minimum mean distance, the maximum and minimum standard deviations resulting for each type of network, for a few characteristics τ values are given when Th_q is calculated for $q = 1.8315$. Recall that the "mean" is that of the distances between nodes on the indicated network in the ensemble of networks generated for the given time windows. The values of the averaging windows T_1 and T_2 when this maximum (minimum) occurs are indicated; the corresponding number of networks (N_{Net}) used for the statistics is also indicated.
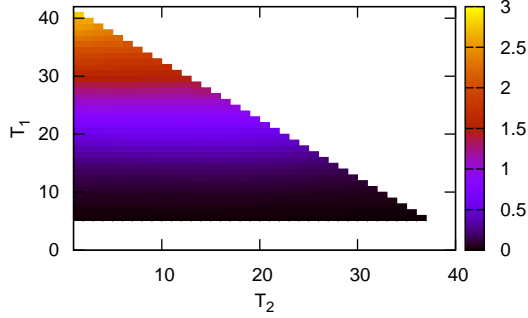

FIG. 3: Mean distance between countries deduced from $Th_{\mathbf{q}}$ for UMLP networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 10$ yrs. The color scale in use is indicated. A few numerical values are given in Table I.
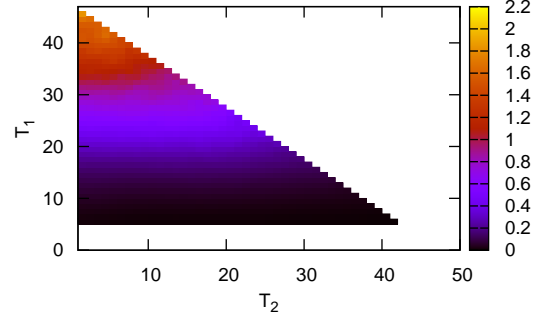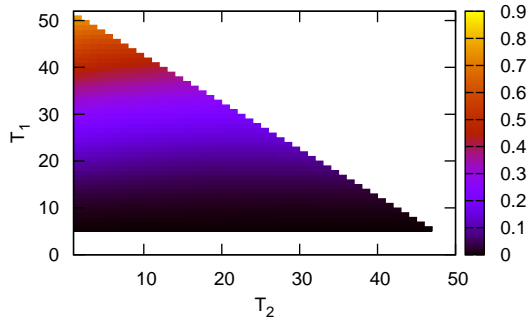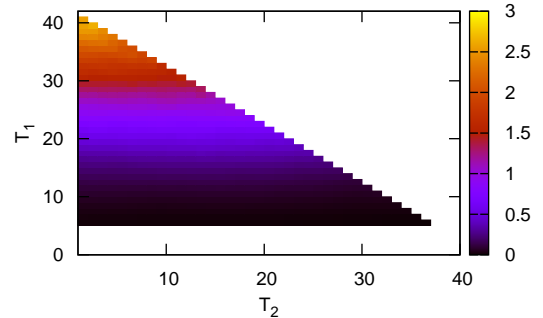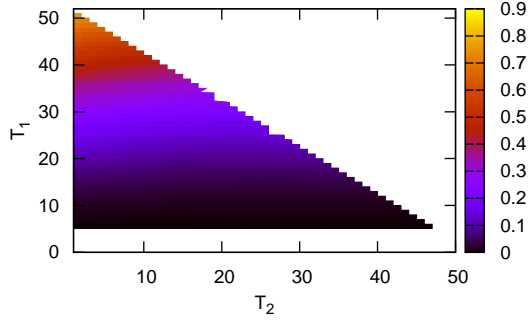

FIG. 5: Mean distance between countries deduced from $Th_{\mathbf{q}}$ for BMLP networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 5$ yrs. The color scale in use is indicated. A few numerical values are given in Table I.
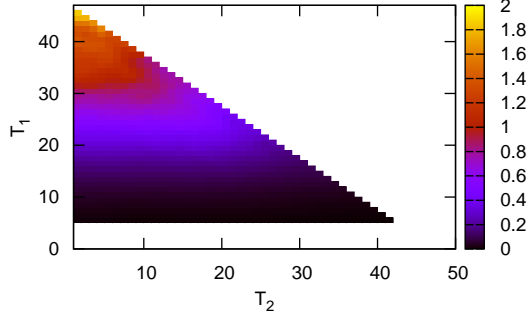

FIG. 4: Mean distance between countries deduced from $Th_{\mathbf{q}}$ for BMLP networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 0$ y. The color scale in use is indicated. A few numerical values are given in Table I.
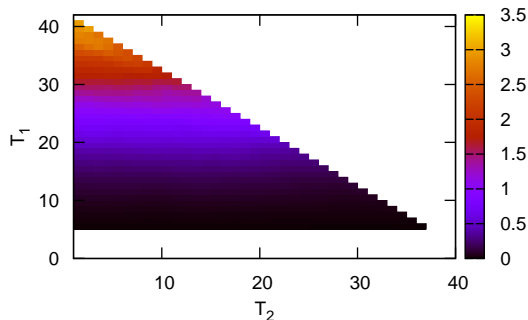

FIG. 6: Mean distance between countries deduced from $Th_{\mathbf{q}}$ for BMLP networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 10$ yrs. The color scale in use is indicated. A few numerical values are given in Table I.

FIG. 7: Mean distance between countries deduced from Th_q for LMST networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 0$ y. The color scale in use is indicated. A few numerical values are given in Table I.

FIG. 8: Mean distance between countries deduced from Th_q for LMST network. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 5$ yrs. The color scale in use is indicated. A few numerical values are given in Table I.

FIG. 9: Mean distance between countries deduced from Th_q for LMST networks. The distance is averaged over the network links and the time. The size of T_1 and T_2 are presented on the vertical and horizontal axis respectively. Time lag $\tau = 0$ yrs. The color scale in use is indicated. A few numerical values are given in Table I.
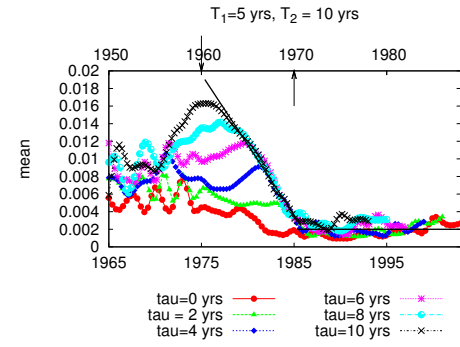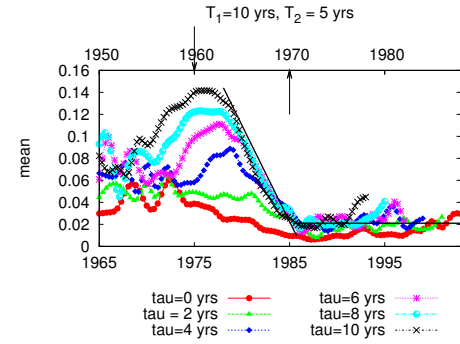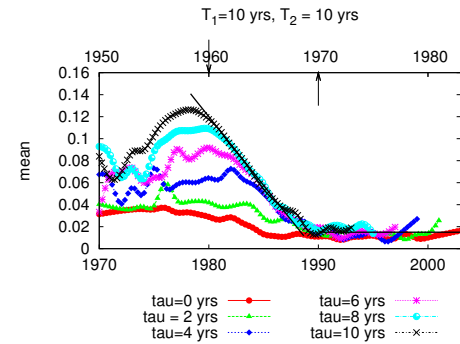

FIG. 10: Yearly evolution of the mean and standard deviation of the links between nodes for the UMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 5$ yrs, $T_2 = 10$ yrs for different time lags τ.
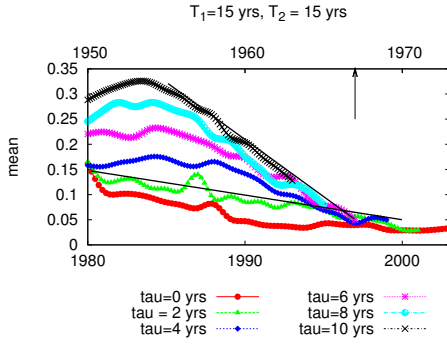

FIG. 11: Yearly evolution of the mean and standard deviation of the links between nodes for the UMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 10$ yrs, $T_2 = 5$ yrs for different time lags τ.

4.2. *q-Theil* network evolution

For further discussion the following time window sizes were chosen, i.e. $(T_1 = 5$ yrs, $T_2 = 10$ yrs$)$, $(T_1 = 10$ yrs, $T_2 = 5$ yrs$)$, $(T_1 = 10$ yrs, $T_2 = 10$ yrs$)$, $(T_1 = 15$ yrs, $T_2 = 15$ yrs$)$, for the three time lags $\tau = 0, 5, 10$ yrs. The evolutions of
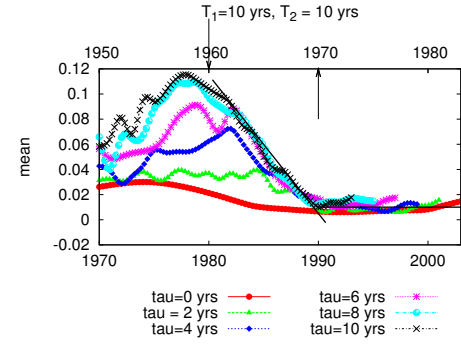

FIG. 12: Yearly evolution of the mean and standard deviation of the links between nodes for the UMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 10$ yrs, $T_2 = 10$ yrs for different time lags τ.
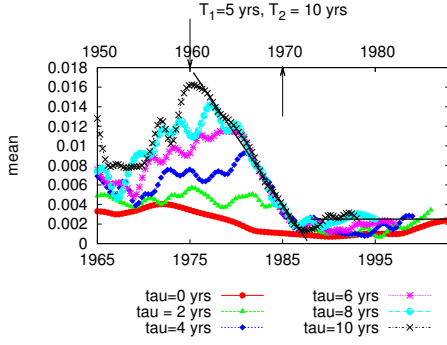
FIG. 13: Yearly evolution of the mean and standard deviation of the links between nodes for the UMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 15$ yrs, $T_2 = 15$ yrs for different time lags τ.

FIG. 16: Yearly evolution of the mean and standard deviation of the links between nodes for the BMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 10$ yrs, $T_2 = 10$ yrs for different time lags τ.
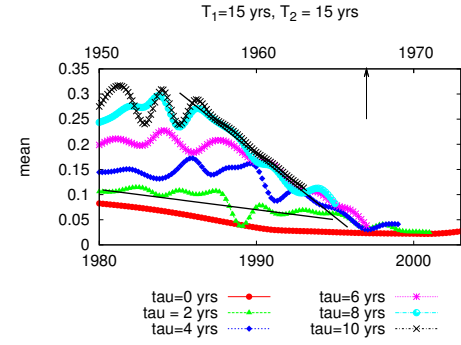

FIG. 14: Yearly evolution of the mean and standard deviation of the links between nodes for the BMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 5$ yrs, $T_2 = 10$ yrs for different time lags τ.

FIG. 17: Yearly evolution of the mean and standard deviation of the links between nodes for the BMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 15$ yrs, $T_2 = 15$ yrs for different time lags τ.

mean distance between countries and the corresponding standard deviations for these chosen time windows sizes are presented in Figs. 10-21. Arrows and straight lines indicate remarkable features.

The general observations to be made at this stage are the following

- In all considered networks (UMLP, BMLP and LMST) and for all window sizes three types of evolution can be distinguished: increase, decrease and relatively stable mean distance between countries.

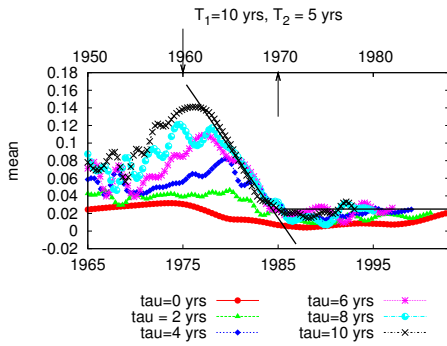- These three types of evolution are better seen for long

FIG. 15: Yearly evolution of the mean and standard deviation of the links between nodes for the BMLP network deduced from Th_q analysis of GDP countries, when $T_1 = 10$ yrs, $T_2 = 5$ yrs for different time lags τ.
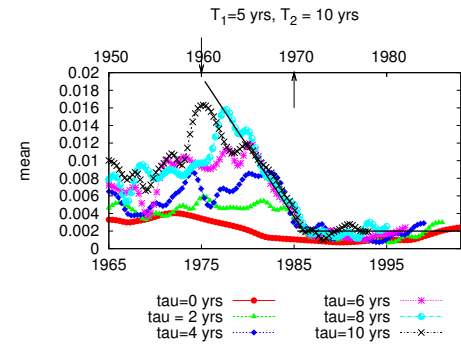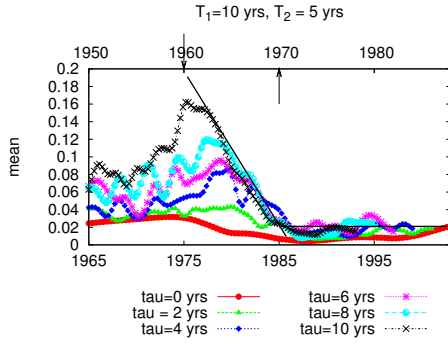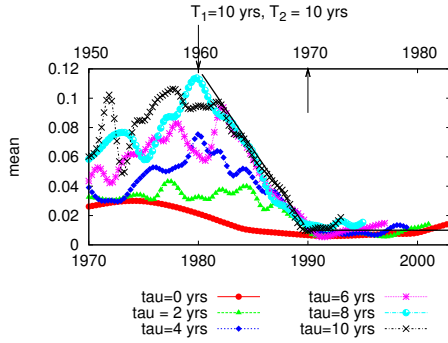

FIG. 18: Yearly evolution of the mean and standard deviation of the links between nodes for the LMST network deduced from Th_q analysis of GDP countries, when $T_1 = 5$ yrs, $T_2 = 10$ yrs for different time lags τ.

FIG. 19: Yearly evolution of the mean and standard deviation of the links between nodes for the LMST network deduced from $Th_\mathbf{q}$ analysis of GDP countries, when $T_1 = 10$ yrs, $T_2 = 5$ yrs for different time lags τ.

FIG. 20: Yearly evolution of the mean and standard deviation of the links between nodes for the LMST network deduced from $Th_\mathbf{q}$ analysis of GDP countries, when $T_1 = 10$ yrs, $T_2 = 10$ yrs for different time lags τ.

lag time ($\tau > 5$ yrs). Therefore the lag time seems to be crucial in any analysis and discussion of the globalization process. This might suggest that some countries play a role of leaders while other follow their way.

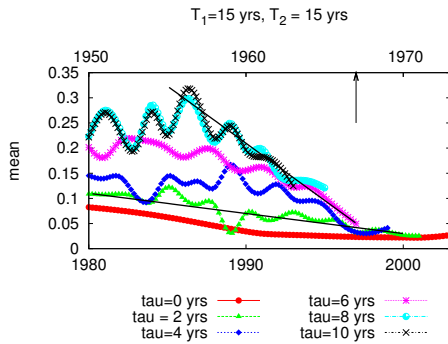- It is worth noticing that for the very long lag time $\tau =$

FIG. 21: Yearly evolution of the mean and standard deviation of the links between nodes for the LMST network deduced from $Th_\mathbf{q}$ analysis of GDP countries, when $T_1 = 15$ yrs, $T_2 = 15$ yrs for different time lags τ.

10 yrs and time windows [($T_1 = 5$ yrs, $T_2 = 10$ yrs), ($T_1 = 10$ yrs, $T_2 = 5$yrs), ($T_1 = 10$yrs, $T_= 10$yrs)] the maximum of the mean distance occurs about 1960, and

- since then the size of the network(s) is fast decreasing over a decade up to 1970 and

- thereafter remains small and relatively stable up to 2000 or so

- when the mean size seems to reincrease.

5. CONCLUSIONS

In conclusion, the most interesting results of this analysis are

- The analysis shows the existence of a globalization process since 1960 till 1970 and its stabilisation thereafter, followed by a destabilisation after 2000 as observed in the decrease of the network size.

- The observation of the globalization process does not depend on the type of network constructed.

- The mean distance between countries and the corresponding std are the largest for the UMLP networks and the smallest for the corresponding LMST networks.

- With increasing time lag the Theil mapping window size T_1 at which the maximum of the network size is found is always decreasing.

- The globalization process is better seen if the lag time is greater than 5 yrs, - which might be considered as the time needed for some synchronization process, but is also in fact commensurate with most government life times and election time intervals. These conjectures suggest further investigations.

- Even though for large time lags the mean values are large, the globalization evolution is the same as for short time lags (greater than 5 yrs); thus a large time lag magnifies the globalization process feature which is easilier to observe then.

Of course much more work is in order to connect the above to some non extensive thermostatistics ideas. To search for a robust ("optimal") q value and the significance of the q-*Theil* index are open questions. Finally let us stress the interest of studying graphs, in particular to derive weighted networks such as in this paper, in order to have some comparative data organisation coherence.

5.1. Acknowldgements

congratulated for his great insight, what he brought to modern equilibrium thermostatistics and for suggesting MA participa-

tion at such a meeting.

---

[1] L. Boltzmann, *Vorlesungen über Gastheorie: 2 Volumes* (Leipzig, 1895/98).

[2] S. Gheorghiu-Svirschevski, Phys. Rev. A **63**, 022105 (2001).

[3] C. E. Shannon, The Bell System Technical Journal **30**, 50 (1950).

[4] J. L. Ruiz de la Torre, A. Velarde, and X. Manteca, Animal Behaviour **59**, 269 (2000).

[5] D. R. Brooks and E. O. Wiley, *Evolution As Entropy: Toward a Unified Theory of Biology* (University of Chicago Press, 1988).

[6] C. T. Murray Gell-Mann, *Nonextensive Entropy: Interdisciplinary Applications* (Oxford University Press, 2004).

[7] M. Ausloos and K. Ivanova, Phys. Rev. E **68**, 046122 (2003).

[8] J. A. Duro and J. Esteban, Economics Letters **60**, 269 (1998).

[9] J. A. James and M. Thomas, Journal of Income Distribution **9**, 39 (2000).

[10] C. Beck and E. G. D. Cohen, Physica A **322**, 267 (2003).

[11] A. Mathai and H. Haubold, Physica A **385**, 493 (2007).

[12] H. Touchette and C. Beck, Phys. Rev. E **71**, 016131 (2005).

[13] C. Tsallis and A. M. C. Souza, Phys. Rev. E **67**, 026106 (2003).

[14] J.-P. Bouchaud and M. Potters, *Theory of Financial Risk and Derivative Pricing* (Cambridge University Press, 2003).

[15] C. Beck, E. G. D. Cohen, and H. L. Swinney, Phys. Rev. E **72**, 056133 (2005).

[16] *H. Theil was a Dutch econometrician who was born on13 October 1924 in Amsterdam, graduated from the University of Amsterdam, succeeded to Jan Tinbergen at the Erasmus University Rotterdam, moved and taught later in Chicago and at the University of Florida. He died in 2000.*

[17] J. Miśkiewicz, Physica A **387**, 6595 (2008).

[18] J. Miśkiewicz and M. Ausloos, Acta Phys. Pol. B **36**, 2477 (2005).

[19] J. Miśkiewicz and M. Ausloos, Int. J. Mod. Phys. C **17**, 317 (2006).

[20] J. Miśkiewicz and M. Ausloos, Physica A **387**, 6584 (2008).

[21] M. Ausloos and R. Lambiotte, Physica A **382**, 16 (2007).

[22] M. Ausloos and M. Gligor, Eur. Phys. J B **57**, 139 (2007).

[23] M. Gligor and M. Ausloos, J. Econ. Integration **23**, 297 (2008).

[24] M. Ausloos and M. Gligor, Acta Phys. Polon. A **114**, 491 (2008).

[25] M. Gligor and M. Ausloos, Eur. Phys. J. B **63**, 533 (2008).

[26] A. C. C. Shalizi and M. Newman, *Power-law distributions in empirical data*, E-print, arXiv:0706.1062 (2007).

[27] E. P. Borges, Phys. A **334**, 255 (2004).